Volume 4, Issue 7, July 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Keyword Tree Analysis Improved PSO Approach for News Article Classification

Dikshi
Pursuing M.Tech
CBS Group of Institution
Jhajjar, Haryana, India

Rahul Kadian
Assistant Professor
CBS Group of Institution
Jhajjar, Haryana, India

Abstract: ***To process on a large group of news articles, research abstract or some other textual information is quite thorny. To read and categorize these documents manually is very time consuming process that even not ensure the accuracy because of diverse contents. To represent these documents effectively, there is the requirement of some document classification approach. In this present work, an optimized hybrid algorithm is defined to present the document clustering and classification. The presented work is divided in three main stages. In first stage, the pre-processing over the individual documents is performed to identify the document keywords. The keyword extraction will perform the word similarity analysis and the word frequency analysis. Once the keywords are extracted, this keyword set will present the document itself. In second stage, the tree based hybrid clustering will be performed based on these keywords set. The hybridization will be here obtained in terms of tree generation and identification of local best and global best using PSO approach. Based on this hybrid approach, the document decision words will be identified. At the final stage, the similarity analysis under different vectors will be performed to keep the similar documents in one group. The similarity measure that will be used in this work are FMeasure and Entropy analysis. The presented work will be implemented in java environment and the analysis of the work will be done under different parameters.***

Keywords: Clustering analysis, document classification, hybrid approach, PSO approach

I. INTRODUCTION

CLUSTERING ANALYSIS

Clustering analysis is a very important technique used in the field of Data Mining. It is a process of grouping together similar multi dimensional data vectors into a number of clusters. The main objective of clustering is to identify and extract important patterns in the underlying data so as to play down inter-cluster similarity and to capitalize on intra-cluster similarity. Clustering Analysis is an ‘unsupervised’ classification where we do not know the class labels. It has found its application in many areas like exploratory data analysis, image segmentation, web mining, mathematical programming, document organization, customer segmentation in marketing etc. Clustering techniques are basically divided into the following types:

- *Hierarchical clustering*: This approach provides a series of nested partitions of the dataset. It divides the data into a nested tree structure where the levels of the tree show similarity or dissimilarity among the clusters at different levels.

- *Partitioning Clustering*: In contrast to hierarchical technique which yield a successive level of clusters by iterative fusions or divisions, this technique assigns a set of objects to clusters with no hierarchical structure. These methods try to minimize certain criteria, like square error function.

- *Distance based Clustering Approach*: This clustering is based on density (local cluster criterion), such as density-connected points. They can handle noise and require a single scan e.g. DBSCAN, OPTICS, CLIQUE, DENCLUE.

- *Grid–Based Clustering Approach:* clustering is done using a multi-resolution grid data structure. Examples of this method are STING (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997), BANG-clustering, GRIDCLUS (Grid-Clustering) by Schikuta (1997),WaveCluster (a multi-resolution clustering approach using wavelet method) by Sheikholeslami, Chatterjee and Zhang (1998).

- *Model based Clustering Methods*: Use certain models for clusters and attempt to optimize the fit between the data and the model

DOCUMENT Classification

With the increase in amount of information, document Classification is applied to the information for easy recognition of the relevance of content. Document Classification is an automatic grouping of text documents into clusters so that documents within a cluster have high similarity in comparison to one another, but are dissimilar to documents in other clusters. Unlike document classification no labelled documents are provided in Classification; hence, document Classification is a kind of unsupervised learning. Document Classification is widely applicable in areas such as search engines, web mining, information retrieval and topological analysis. Its usefulness comes into sight when it reduces the search space required to respond to a query. It has become an increasingly important task in analyzing huge numbers of documents distributed among various sites. The demanding aspect is to analyze this massive number of extremely high dimensional distributed documents and to organize them in such a way that results in better search and knowledge extraction without introducing much extra cost and complexity. Most of the algorithms for document Classification can be classified as either hierarchical or partitioning methods but in it has been proved that partitioning algorithms perform better than hierarchical algorithms.

II. LITERATURE REVIEW

In Year 2005, JIAN-SUO XU performed a work,**” TCBLHT: A NEW METHOD OF HIERARCHICAL TEXT CLUSTERING”.** This paper presents a new method of hierarchical text clustering based on combination of latent semantic analysis (LSA) and hierarchical TGSOM, which is called TCBLHT method. The TCBLHT method can automatically achieve hierarchical text clustering, and establishes vector space model (VSM) of term Authoright by using the theory of LSA

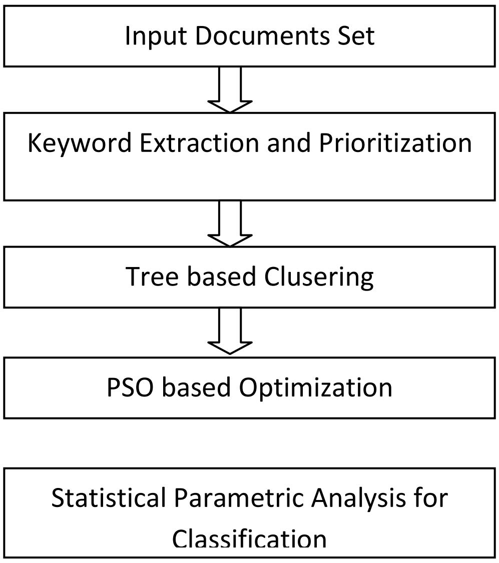
In Year 2005, MAO-TING GAO performed a work,” **RPCL TEXT CLUSTERING BASED ON CONCEPT INDEXING”.** This paper presents a new method using CI to reduce dimensionality for RPCL text clustering. Experimental results show that this algorithm not only improves clustering efficiency greatly, but also makes the averaged accuracy reach a high level.

In Year 2010, Jiabin Deng performed a work,**” An Improved Fuzzy Clustering Method for Text Mining”.**This paper proposes an improved fuzzy clustering-text clustering method based on the fuzzy C-means clustering algorithm and the edit distance algorithm. Author use the feature evaluation to reduce the dimensionality of high-dimensional text vector. Because the clustering results of the traditional fuzzy C-means clustering algorithm lack the stability,

In Year 2007, Yun Sha performed a work,**” Text Clustering Algorithm Based on Lexical Graph”.** An algorithm for text clustering based on lexical graph is proposed in this paper, which is a kind of term-based cluster method. The lexical graph is build with nodes representing words and edges representing their concurrent in text. The attribute of each node is text which the word occurs in. A cluster center is defined as node (word) with large degree in this graph, the center attributes (text occurs in) and its neighbors' are partitioned to one cluster whose description is the center node. This approach reduces drastically the dimensionality of the data and improves the synonymy extension ability.

In Year 2008, Authori Wang performed a work,**” Fuzzy C-Means Text Clustering with Supervised Feature Selection”.** In this paper Author propose a new text clustering algorithm SFFCM which use the supervised feature selection method to select the feature. The SFFCM is based on the EM algorithm. In the E-step, to calculate the expectation, Author use the supervised feature selection algorithm to calculate the relevancy score for each term. In the M step Author use the FCM algorithm to obtain the cluster results based on the selected terms.

In Year 2008, Fuzhi Zhang performed a work,**” An Ant-based Fast Text Clustering Approach Using Pheromone”.** In this approach, however, the ant's moving is random, which leads to the convergence speed too slow. Aims at abovementioned problem, an ant-based fast text clustering approach(AFTC) is presented. This approach utilizes pheromone left by ants to avoid ant's moving randomicity, which can make the ant move towards direction which has high pheromone concentration at each step, and the direction of moving is the orientation where the text vectors are relatively concentration.


Input Documents Set
Keyword Extraction and Prioritization
Tree based Clusering
PSO based Optimization
Statistical Parametric Analysis for Classification

In Year 2009, *Yi Guo* performed a work**," A Hierarchical Text Clustering Algorithm with Cognitive Situation Dimensions".** This paper introduces an innovative research effort, CogHTC, a hierarchical text clustering algorithm, inspired by cognitive situation models. CogHTC extracts representative features from four elaborately selected cognitive situation dimensions with consideration of the clustering efficiency.

Proposed Work

The presented work is about to perform the document categorization based on three layers given as in figure

The proposed work is about to categorize the available set of documents in defined groups. In the first layer, the keyword based analysis is performed and it will use a approach for keyword prioritization. Once the prioritization is done, in second layer the keyword tree analysis improved PSO approach will be defined for document clustering and finally the statistical measures will be applied to perform the document classification.

III. CONCLUSION

The presented work is the implementation of PSO optimized clustering algorithm is defined to cluster the documents. In this work, the clustering is performed on text documents. This clustering is performed at two different level. In the earlier stage of clustering, the structural analysis is performed on text documents. Once the structure is extracted, the next work is to perform content based filteration and mining to group the related data. In this presented work the improvement to the existing clustering approach is done by performing the re-performing the distance based check between the clusters present in hierarchal order. If the hierarchal order is under the specific limit, the child cluster will shift to the parent and aggrative cluster will be composed. In this improved algorithm, the improvement on clustering concept is obtained. The obtained result shows that the presented work is more effective.

IV. FUTURE WORK

I n this present work, an improvement to optimized algorithm is defined to perform the effective clustering based on the structural and content match on TEXT data. The work is defined for text document. The work can be improvrd in future for the following direction:

- The work is performed on text document in future the same work can be defined for some other document type.

REFERENCES

[1] JIAN-SUO XU," **TCBLHT: A NEW METHOD OF HIERARCHICAL TEXT CLUSTERING",** Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou 0-7803-9091-1/05©2005 IEEE

[2] MAO-TING GAO," **RPCL TEXT CLUSTERING BASED ON CONCEPT INDEXING",** Proceedings of the Fourth International Conference on Machine Learning and Cybernetics 0-7803-9091-1/05©2005 IEEE

[3] Jiabin Deng**," An Improved Fuzzy Clustering Method for Text Mining",** 2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing 978-0-7695-4011-5/10© 2010 IEEE

[4] MAO-TING GAO," **A NEW ALGORITHM FOR TEXT CLUSTERING BASED ON PROJECTION PURSUIT",** Proceedings of the Sixth International Conference on Machine Learning and Cybernetics -4244-0973-X/07©2007 IEEE

[5] Yun Sha**," Text Clustering Algorithm Based on Lexical Graph",** Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007) 0-7695-2874-0/07© 2007 IEEE

[6] Wei Wang**," Fuzzy C-Means Text Clustering with Supervised Feature Selection",** Fifth International Conference on Fuzzy Systems and Knowledge Discovery 978-0-7695-3305-6/08© 2008 IEEE

[7] Fuzhi Zhang**," An Ant-based Fast Text Clustering Approach Using Pheromone",** Fifth International Conference on Fuzzy Systems and Knowledge Discovery 978-0-7695-3305-6/08© 2008 IEEE

[8] *Yi Guo*," **A Hierarchical Text Clustering Algorithm with Cognitive Situation Dimensions",** Second International Workshop on Knowledge Discovery and Data Mining 978-0-7695-3543-2/09© 2009 IEEE

[9] Yanping Lu**," Text Clustering via Particle Swarm Optimization",** 978-1-4244-2762-8/09©2009 IEEE

[10] Guo Wensheng**," Text Clustering Algorithm Based on Spectral Graph Seriation",** 978-1-4244-2723-9/09@ 2009 IEEE